A Survey on Text Mining Process and E-Mail Categorization Methods

Minoti Patel[1] Pooja Bhatt[2]
[1,2]Department of Computer Engineering
[1,2]Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430

Abstract— Data mining is the process of extraction of hidden predictive information from the huge databases. It is a new technology with great latent to help companies focus on the most important information in their data warehouses. Text mining is a technique which is used to extract interesting information or knowledge from the text documents. Web mining helps in extraction and integration of useful information and knowledge from webpage contents. Web mining use data mining techniques to automatically discover and extract knowledge and useful information from web document or services. E-mail has been an efficient and popular communication mechanism as the number of internet user's increase. Large quantity of emails is difficult for users to efficiently organize and retrieve. For this purpose, various algorithm and technique are used for categorizations of E-mails. This paper basically focuses on study of the Text mining process and its framework then discusses various techniques for categorizations of Emails.

Key words: Text mining terminology, Tex mining process, Data mining techniques

I. INTRODUCTION

At the present time, every person involve with internet. Everyone has at least one email account for delivering files and important information's with others. Electronic mail is a fast, efficient, low-cost and one of the most preferred way of communication among a people at the same time. Electronic mail can be viewed as a special type of documents along with some identification information like "to", "from", "cc", "subject", "attachments" and so on. User find themselves expend large amount of time and effort sifting through the group of mail messages and classifying them to their equivalent folders. Therefore email management is a vital problem for organization and individuals because it is prone to misuse. One aspect of email management is to classify email messages into appropriate folders automatically [1].

Many researchers widely. There was lots of work done using different machine learning methods for classifying mails into spam and non-spam. This area mainly works on different features and combination of the features. Email classification and categorization is done based on header and body fields as well like from, to, cc, subject, attachments etc. E-Mail categorization is part of text mining area. Text mining also known as text data mining or knowledge discovery from textual databases refers the process of extracting interesting and non-trivial patterns or knowledge from text documents. There are different parts of text mining. One part of text mining is web mining. Web mining is a data mining techniques which automatically discover information from web documents. Web mining is separated in three different types, which are web usage mining, web content mining and web structure mining [1].

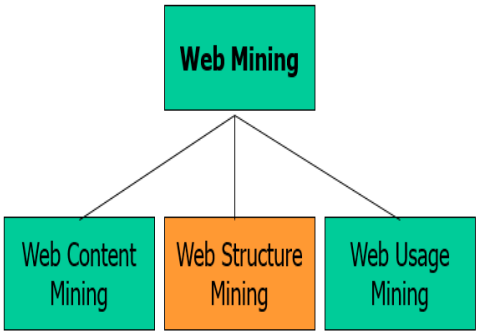

Fig. 1: Common Structure of web mining

II. TEXT MINING TERMINOLOGY

A. *Text mining Vs. Data mining*

In text mining, different patterns are extracted from natural language but in Data mining patterns are extracted from databases.

B. *Text Mining Vs. Web Mining*

In Text mining, applied input is free unstructured text, but in web mining input which is web sources are structured [2].

C. *Text Mining Process*

Text mining process divided in different steps. These all steps of text mining label in figure 2[3].

1) Text Preprocessing

This step further divided into number of following sub steps.

- Tokenization: Text documents contain various sentences. In this step divide whole sentence into words and then removing space, commas etc.
- Stop word Removal: These steps involve removing of HTML, XML tags from web pages.
  Then do the process of removing stop words.
- Stemming: This technique is used to find out root/stem of a word. Stemming converts words in their stream.

2) Text transformation or Feature Generation

Text transformation or Feature generation means convert text documents in words, which can be used for effective analysis task.

3) Feature selection or Attribute selection

This phase mainly used for remove irrelevant features. These procedures give benefit like smaller dataset size, less computation and minimum search space.

4) Text mining methods

Data mining contain different text mining methods likes Topic extraction, Clustering, Classification, Information retrieval, topic discovery and summarization.

5) Evaluation

This phase include evaluation of result in terms of calculating Precision, Recall, Accuracy, F measure etc.

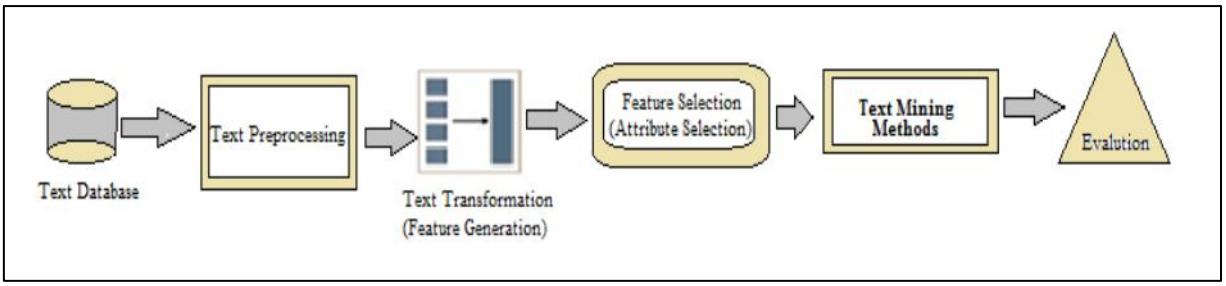

Fig. 2: General Text mining flow

III. General Challenges

Some different challenges are:

- Each user's has different mentality for categorizing emails which they done manually hence criteria used may not be same for all users. This
- Each user's mail box is different and that mailbox changing every day. Folder content also varies because new messages are added and old messages are deleted.
- Categorize emails to appropriate sub folders will require a correct knowledge understanding characteristics of each folders.
- Information overload: every day lots of user create new account so sorting this data and separating those goods from the bad is a more difficult task.
- Message handling: message handling time consuming because currently message handle manually.
- Protection: user may have different account or email addresses for their different requirements. In addition over time user may change email accounts. Currently there are no system to profile the user on one account and apply this profile to a new or shared account to provide fraud and protection of misuse.
- Folder characteristic may vary for more number of emails means folder size varies from time to time. So classification system needs well performance even if large training set is absent [1].

IV. Data Mining Techniques

In this section, various data mining techniques which is used for categorizing emails is discussed via different research groups.

A. Support Vector Machine(SVM)

SVM is a supervised learning method which is used for emails categorization.SVM also known as binary classifier. SVM classifiers are binary classifier so it is used for spam detection or classify emails in only two classes. so if any user want to classify emails in spam or non spam or only two classes then this method is used for email categorization. SVM method gives better performance than Naïve bayes classifier. The main goal of SVM is to find a hyperplane. This hyperplane maximize the margine between two classes[4].

1) Advantages

- SVM provide good sample of generalization.
- SVM deliver a uniqe solution if problem is convex.

2) Disadvantages

- Lies in choice of the kernal
- Discrete data presentation

B. Naïve Bayes Classifier

Naïve Bayes Classifier widely used for detection of spam emails. This method assume that features values are statistically autonomous from each other. This method depends on probablistic relationship between different categories. Bayes theorem combine all probabilities of interesting features. If probability is closer to 0 then message is non spam and if probability closer to 1 then message is spam[5].

1) Advantages

- Naïve Bayes is not support irrelevent features.
- Handles discreate and real data

2) Disadvantages

- Independence features is assumes.

C. Vector Space Model(VSM)

Now a day's Email categorization is changing due to sparse feature space. For this problem vector space model is used. Basic idea of VSM is to take related semantic feature and use that features to enrich the semantic feature of an email and do categorization. Experimental evaluation shows VSM produce better accuracy for smaller training sets. This method produces better result than other categorization method [6].

1) Advantages

- Simple computational framework for ranking
- Any similarity measure could be used

2) Disadvantages

- Assumption of term independent
- For effective ranking there is no prediction about techniques

D. Hidden Morkov Model (HMM)

Hidden markov model is a probabilistic model which is used for model time series data. HMM is statistical tool which is used for model generative sequences for the Process in which observation is of probabilistic function of the states. In this, hidden states are not directly visible [1].

1) Advantages

- Handle record structure variation

2) Disadvantages

- Require training for annotated data
- Manual mark-up may required

E. Conditional Random Field (CRF)

Conditional random field is undirected graphical model. Conditional random field is a extension of maximum entropy. CRF consider past and future dependency for finding transition probability. CRF calculate conditional probability for output vertices based on input [7].

1) Advantages
- Multiple interacting features is allowed by conditional random field.

2) Disadvantages
- Slow convergence during training.

V. CONCLUSION

This paper presents the overview of email and also discusses the overview of text mining process. In text mining process different steps of text mining discuss which shows how text mining process actually works. Different terminology of text mining is discussed. Different methods discussed here are capable for categorizing emails. Advantages and disadvantages of method is display.

REFERENCES

[1] Shreyansh U. Saraiya, Prof. Nikita Desai, "Content Based Categorization of E-Mail using Hidden Markov Model Approach" in IEEE 2014
[2] Rahmi Agrawal, Mridula Batra "A Detailed Study on Text Mining Techniques " in IJSCE 2013
[3] Falguni N. Patel, Neha R. Soni "Text mining: A Brief survey "in International Journal of Advanced Computer Research December 2012
[4] Upasana, S.Chakravarty "A Survey of Text Classification Techniques for E-mail Filtering" in IEEE 2010
[5] Amam Dixit, Anjura Arora "Text and Image Based Spam Email Classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm in IEEE 2014.
[6] http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html
[7] http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/

